Hu et al. BMC Bioinformatics 2011, 12:364
http://www.biomedcentral.com/1471-2105/12/364

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Characterizing genetic interactions in human disease association studies using statistical epistasis networks

Ting Hu[1], Nicholas A Sinnott-Armstrong[1], Jeff W Kiralis[1], Angeline S Andrew[2,3], Margaret R Karagas[2,3] and Jason H Moore[1,2,3*]

Abstract

Background: Epistasis is recognized ubiquitous in the genetic architecture of complex traits such as disease susceptibility. Experimental studies in model organisms have revealed extensive evidence of biological interactions among genes. Meanwhile, statistical and computational studies in human populations have suggested non-additive effects of genetic variation on complex traits. Although these studies form a baseline for understanding the genetic architecture of complex traits, to date they have only considered interactions among a small number of genetic variants. Our goal here is to use network science to determine the extent to which non-additive interactions exist beyond small subsets of genetic variants. We infer statistical epistasis networks to characterize the global space of pairwise interactions among approximately 1500 Single Nucleotide Polymorphisms (SNPs) spanning nearly 500 cancer susceptibility genes in a large population-based study of bladder cancer.

Results: The statistical epistasis network was built by linking pairs of SNPs if their pairwise interactions were stronger than a systematically derived threshold. Its topology clearly differentiated this real-data network from networks obtained from permutations of the same data under the null hypothesis that no association exists between genotype and phenotype. The network had a significantly higher number of hub SNPs and, interestingly, these hub SNPs were not necessarily with high main effects. The network had a largest connected component of 39 SNPs that was absent in any other permuted-data networks. In addition, the vertex degrees of this network were distinctively found following an approximate power-law distribution and its topology appeared scale-free.

Conclusions: In contrast to many existing techniques focusing on high main-effect SNPs or models of several interacting SNPs, our network approach characterized a global picture of gene-gene interactions in a population-based genetic data. The network was built using pairwise interactions, and its distinctive network topology and large connected components indicated joint effects in a large set of SNPs. Our observations suggested that this particular statistical epistasis network captured important features of the genetic architecture of bladder cancer that have not been described previously.

Background

Identifying associations between genetic and phenotypic variation is crucial to understanding the genetic basis of disease susceptibility and disease etiology [1], and to devising diagnostic tests and useful treatments [2,3]. With the rapid expansion of open-access single nucleotide polymorphism (SNP) databases [4], the progress in genotyping technologies [5], and the availability of immense computational resources [6], mapping the genes that underlie common diseases and quantitative traits is now feasible.

Genome-wide associations studies (GWAS), in which thousands to millions of SNPs across the human genome are tested for associations with disease phenotypes, have emerged as a particularly promising approach for drawing causal inferences between traits and genetic variation [2,3,7,8]. However, although GWAS have uncovered numerous disease susceptibility loci [3,8,9],

* Correspondence: jason.h.moore@dartmouth.edu
[1]Department of Genetics, Dartmouth Medical School, Dartmouth College, Lebanon, NH, USA
Full list of author information is available at the end of the article

the majority of them have had relatively subtle individual associations with disease risk. The success of GWAS analyzed only for individual SNP effects largely depends on fundamental assumptions about a lack of genetic complexity and a simple single-gene architecture of diseases, and becomes greatly compromised when gene-environment or gene-gene interactions modify the relationship between genotypes and phenotypes [10-13].

The genetic architecture of common human diseases is, in fact, characterized in part by interactions between genes, i.e., *epistasis* [13-19]. Accordingly, the focus of recent research has shifted from identifying single locus susceptibility [2,7] to quantifying interaction effects between multiple candidate loci throughout the human genome [13,16,20,21]. However, the study of epistasis faces an initial challenge arising from the existence of fundamental differences between the concepts of biological and statistical interaction (e.g. [21]). These differences imply that *statistical epistasis*, defined at the population level as the non-additive mathematical relationship among multiple genetic variants, cannot be literally translated into *biological epistasis*, which is the physical interaction among two or more molecules at the cellular level of an organism, and vice-versa [17]. Moreover, detecting gene-gene interactions and accounting for them in GWAS further represents a statistical and computational challenge [12,13,20,22]. The statistical challenge results from the prohibitive amount of data necessary to support the huge number of hypotheses involved in modeling interactions, even when considering only pairwise interactions [3,11]. The computational challenge, in turn, arises from the necessity to exhaustively evaluate all possible combinations of SNPs, which becomes infeasible when interactions involve more than two SNPs: the computational complexity, which is in the quadratic order for pairwise interactions, increases exponentially with higher-order interactions, rendering any exhaustive assessment impossible [12,13,21].

The necessity to overcome these difficulties calls for efficient tools to detect genetic interactions [2,7,23]. Methods such as machine learning [24-26] and dimensionality reduction [27,28] have recently proven useful in detecting influential interactions. However, these approaches are aimed at identifying best models consisting of several SNPs and thus ignore the broader gene-gene interaction landscape.

A particularly intuitive approach to explore the genetic architecture of common human diseases and to identify genetic interactions is to use networks. A network is generally defined as a collection of vertices joined in pairs by edges and is a powerful tool to represent and study complex systems [29,30]. In biological systems, for instance, networks can be used to characterize interactions at all levels of organization, from the molecular level with metabolic [31,32], protein-protein interaction [33], and genetic regulatory networks [34], to the macroscopic level with food webs [35].

Networks allow for a structured representation of a collection of entities and their relationships, which provides a well-suited framework for the study of epistasis. The use of networks does not resolve the dimensionality problems inherent in exploring high-order interactions amongst multiple SNPs. An intuitive solution that has previously proven helpful is to filter out the considerable noise masking the useful genotypes and to reduce the search space to a subset of high-susceptibility SNPs before constructing a network of genetic interactions.

An example of such a sequential approach is the work of McKinney et al. [36], who developed a genetic-association interaction network to visualize and interpret synergetic interactions between pairs of SNPs. Loci were initially chosen based on the strength of their main effects. Although useful, purging databases for irrelevant genetic variants and preliminarily selecting high-susceptibility SNPs inevitably comes at the risk of discarding loci comprised in significant higher order interactions. Hence, alternative solutions for reducing the space of possible interactions in GWAS are needed.

In the present study, we propose to infer genetic interaction networks that are not dependent on statistical main effects. We first rank all possible pairwise interactions between SNPs according to their relative strength and subsequently build and analyze *statistical epistasis networks* including only those interactions whose strength exceeds a given threshold. Hence, the approach we apply distinguishes itself from existing ones in the following ways: 1) We qualify the strength of all pairwise interactions identifiable in the complete data set rather than a subset of high main-effect SNPs; 2) We organize our genetic network around the strongest pairwise interactions rather than around the strongest main effects; 3) We analyze network topologies to systematically identify the network that best captures the genetic architecture inherent in the data; 4) In contrast to many existing techniques that aim at identifying a classification model consisting of a subset of susceptibility SNPs, our epistasis network captures a broader landscape of gene-gene interactions through exhaustively enumerating all possible pairwise interactions.

In the United States, bladder cancer is one of the most common types of cancer in both men and women. Although the main known cause of bladder cancer is smoking [37], recent case-control studies also suggest that there exist heritable susceptibility factors [38-40]. Thus, we used the network approach to characterize the space of pairwise interactions in a bladder cancer data

set consisting of 1,422 SNPs sampled across 491 patients newly diagnosed bladder cancer and 791 controls [41]. Statistical epistasis networks were built by incrementally adding edges between SNPs if the strength of their pairwise interactions was greater than a given threshold. We identified one threshold value for which the resulting network showed unique topological characteristics, which we believe, capture the complex structure intrinsic in the data. Its distinctively large connected component suggests that a group of SNPs may jointly modify the disease outcome. Thus, the network may serve as a scaffold to explore the landscape of higher-order interactions.

Methods

Bladder cancer data set

The data set used in this study consisted of cases of bladder cancer among New Hampshire residents, ages 25 to 74 years, diagnosed from July 1, 1994 to June 30, 2001 and registered in the State Cancer Registry. All controls were selected from population lists. Controls less than 65 years of age were selected using population lists obtained from the New Hampshire Department of Transportation, while controls aged 65 and older were chosen from data files provided by the Centers for Medicare & Medicaid Services (CMS) of New Hampshire. This data set also shared a control group with a study of non-melanoma skin cancer in New Hampshire covering an overlapping diagnostic period of July 1, 1993 to June 30, 1995 and July 1, 1997 to March 30, 2000. Additional controls were selected for bladder cancer cases diagnosed in the intervening period frequency matched to these cases on age (25-34, 35-44, 45-54, 55-64, 65-69, 70-74 years) and gender.

Informed consent was obtained from each participant and all procedures and study materials were approved by the Committee for the Protection of Human Subjects at Dartmouth College. Consenting participants underwent a detailed in-person interview, usually at their homes. Recruitment procedures for both the shared controls from the non-melanoma skin cancer study and additional controls were identical and ongoing concomitantly with the case interviews.

DNA was isolated from peripheral circulating blood lymphocyte specimens harvested at the time of interview using Qiagen genomic DNA extraction kits (QIAGEN Inc., Valencia, CA). Genotyping was performed on all DNA samples of sufficient concentration, using the GoldenGate Assay system by Illumina's Custom Genetic Analysis service (Illumina, Inc., San Diego, CA). Out of the submitted samples, 99.5% were successfully genotyped and samples repeated on multiple plates yielded the same call for 99.9% of SNPs. The missing genotypes were imputed using a frequency-based method. That is, the missing value of an individual was filled using the most common genotype of the corresponding SNP in the population. The data set used in our analysis consisted of 491 bladder cancer cases and 791 controls and most (> 95%) of the subjects were of Caucasian origin. More details on this data set and the methods are available in [40,41].

Network construction

Networks are formalized mathematically by graphs, and we use these two terms interchangeably in this article. A graph G is composed of a set $V(G)$ of vertices and a set $E(G)$ of edges [42]. In our epistasis networks, each vertex corresponds to a SNP, and we use v_A to denote the vertex corresponding to SNP A. An edge linking a pair of vertices, for instance v_A and v_B, corresponds to an interaction between SNPs A and B.

We first assigned a weight to each SNP and each pair of SNPs to quantify how much of the disease status the corresponding SNP and SNP pair genotypes explain. In analogy to statistical models, those weights correspond to the strength of the main and the interaction effects and stronger effects translate into higher weights. In information theoretic terms, those weights correspond to the so-called *mutual information* and *information gain* [43]. Specifically, the weight of SNP A is $I(A; C)$, the mutual information of SNP A's genotype and C, the class variable with status *case* or *control*. Intuitively, $I(A; C)$ is the reduction in the uncertainty of the class C due to knowledge about SNP A's genotype. Its precise definition is

$$I(A; C) = H(C) - H(C|A), \tag{1}$$

where $H(C)$ is the *entropy* of C, i.e., the measure of the uncertainty of class C, and $H(C|A)$ is the *conditional entropy* of C given knowledge of SNP A. Entropy and conditional entropy are defined by

$$H(C) = \sum_{c} p(c) \log \frac{1}{p(c)}, \tag{2}$$

$$H(C|A) = \sum_{a,c} p(a, c) \log \frac{1}{p(c\,|a\,)}, \tag{3}$$

where $p(c)$ is the probability that an individual has class c, $p(a, c)$ is that of having genotype a and class c, and $p(c|a)$ is that of having class c given the occurrence of genotype a. In our implementation, $p(c)$ is the frequency of diseased (case) or healthy (control) individuals respectively, $p(a, c)$ is the frequency of individuals in either the case or the control group that carry genotype a, and $p(c|a) = p(a, c)/p(a)$, where $p(a)$ is the frequency of individuals that have genotype a. Given that in most

cases a SNP has two alleles and there are consequently three possible genotypes for each SNP in the diploid human genome, the sum in equation (3) is over all six possible combinations of genotypes a and classes c. Mutual information $I(A; C)$ takes only non-negative values. If the class C is independent of a SNP A's genotype, $I(A; C) = 0$, i.e., SNP A does not predict the disease status. If a correlation exists between the class C and SNP A, $I(A; C) > 0$, i.e., SNP A has a main effect and predicts some of the disease status. Larger values of $I(A; C)$ indicate stronger correlations between A and C.

Given the pair of vertices v_A and v_B, its weight is the information gain $IG(A; B; C)$, where

$$IG(A; B; C) = I(A, B; C) - I(A; C) - I(B; C). \quad (4)$$

As such, $IG(A; B; C)$ is the reduction in the uncertainty, or the information gained, about the class C from the genotypes of SNPs A and B considered together minus that from each of these SNPs considered separately. In brief, $IG(A; B; C)$ measures the amount of synergetic influence SNPs A and B have on class C. A higher value indicates a stronger synergetic interaction. Note that $IG(A; B; C)$ can take non-positive values. A negative value indicates that the genotypes of two SNPs tend to vary together (redundant information), while a value of zero indicates either that the genotypes of the two SNPs are independent or, more likely, that they interact with a mixture of synergy and redundancy. The synergetic part of the mix tends to make the information gain positive while the redundant part lowers the information gain.

Information theory has previously been applied in epistasis studies. For instance, Moore et al. [44,45] used interaction dendrograms based on information gain to interpret their epistasis models. McKinney et al. [36] used information gain to quantify synergic interactions between pairs of SNP in their genetic-association interaction network. In a more general framework, Jakulin and Bratko [46] used mutual information and information gain to quantify the information shared by single class variables and their corresponding attributes. Although there are many other approaches, such as MDR, random forest, and logistic regression, that are able to measure the strength of main and interaction effects of SNPs, we specifically chose information theoretical measures in this study because they are more computationally efficient than the others. This is very important in the era of GWAS since inferring interactions on a genome-wide scale is very computationally intensive.

We then built a series of statistical epistasis networks by incrementally adding edges. These networks were denoted by G_t, where edges between SNPs were added only if their pair weights were greater than or equal to a threshold t. The threshold t varied between 0 and the maximum pair weight estimated based on the data. The networks G_t grew as the threshold t decreased. For $t_1 < t_2$, G_{t_1} contained all the edges and vertices of G_{t_2}.

Network analysis

Our analysis method relies on comparisons between the real data set and its derivatives generated by permutation testing. First, permuted data were used to assess the significance level of the interaction strength of each SNP pair. Second, and more importantly, by comparing networks built from real data and permuted data, we can determine the statistical significance of the network properties themselves. We repeated the network construction and characterization exactly the same way on both real data and permuted data. Thus, any network features observed in the real data that were not consistent with the distribution of features from the permuted data can be inferred to be due to real genetic associations.

We generated 1,000 permuted data sets by randomly shuffling the disease status of the 1,282 samples 1,000 times. This removed all biological signals from the data. For each permuted data set, we then calculated the weights for all pairs of SNPs and constructed a series of networks using the same thresholds as when we built the real-data networks. Once all the networks were assembled, we first evaluated the significance of each pair of SNPs in the real data set by calculating the fraction of permuted data sets with pair weight greater than that obtained from the real data. Then, we investigated and compared some basic properties of these series of networks.

The four basic properties of a network considered here are the number of edges, the number of vertices, the size of the largest connected component, and the vertex degree distribution. The definitions of these standard graph-theoretic terms [42] are summarized as follows. A *connected component* of a graph is a maximal connected subgraph, and the size of a connected component refers to its number of vertices. A graph H is a subgraph of G if both the vertex set and edge set of H are subsets of those of G. A subgraph is *connected* if any two vertices in it can be joined by a sequence of edges. The *degree* of a vertex v, denoted by $d(v)$, is the number of edges incident with v. The fraction of vertices in a network that have degree d is denoted by $p(d)$. Thus, $p(d)$ can be viewed as the probability that a randomly chosen vertex in the network has degree d. The quantities $p(d)$ make up the *vertex degree distribution* of a network. In the context of epistasis networks, the degree of vertex v_A indicates how many SNPs interact with SNP A, while the clustering of vertices within a connected

component may help narrow the search for informative SNPs likely to jointly modify disease outcome.

Results

Measures of main and interaction effects in the bladder cancer data

As shown in Figure 1-A, most of the 1,422 SNPs had relatively small main effects (*mean* ± *stdev* = 0.00122 ± 0.00125) and a few SNPs had very strong main effects. The highest weight was 0.01551 for SNP *IGF2AS_04* and the second highest weight, which was about half of the highest, was 0.00832 for *LRP5_12*. The distribution of interaction strengths (Figure 1-B) had *mean* ± *stdev* = 0.00235 ± 0.00171. The highest weight was 0.01967, and corresponded to the interaction between SNPs *ESR2_02* and *TERT_25*. Of all $\binom{1422}{2} = 1,010,331$ pairs of SNPs, there were 778 pairs with a weight of zero, and 3,083 with negative weights.

Network investigations

The four topological features of G_t and of the permuted-data networks were investigated. All these features were found to distinguish the structure of G_t from the permuted-data networks. The network $G_{0.013}$ was of special interest by showing the most significant network topologies, and is considered in some detail at the end of this section.

Numbers of edges and vertices

Recall that the existence of an edge linking SNPs A and B in the epistasis network G_t indicates an interaction of strength $IG(A; B; C) \geq t$ between them and the networks G_t grow as t decreases. Accordingly, the numbers of edges and vertices of G_t increased monotonically as t decreased from 0.02 to 0 in increments of 0.001 (Figure 2). Moreover, the networks G_t had overall more edges and vertices than the corresponding permuted-data networks. Statistically significant differences ($p \leq 0.01$ drawn from permutation testing) in the numbers of edges and vertices present were detected for threshold values satisfying $0.018 \geq t \geq 0.009$.

Size of the largest connected components

Figure 3 shows the size of the largest connected component in the network G_t and in the permuted-data networks as t decreased from 0.015 to 0.007. The largest connected component of G_t grew quickly with decreasing t. A dominant connected component (larger than any other connected components) emerged at t = 0.013 and its growth became considerably steeper subsequently. The largest connected components of the permuted-data graphs, on the other hand, did not start growing before lower values of the threshold were reached, resulting in the major increase in growth happening later than in G_t. Accordingly, their sizes were smaller for most values of the threshold.

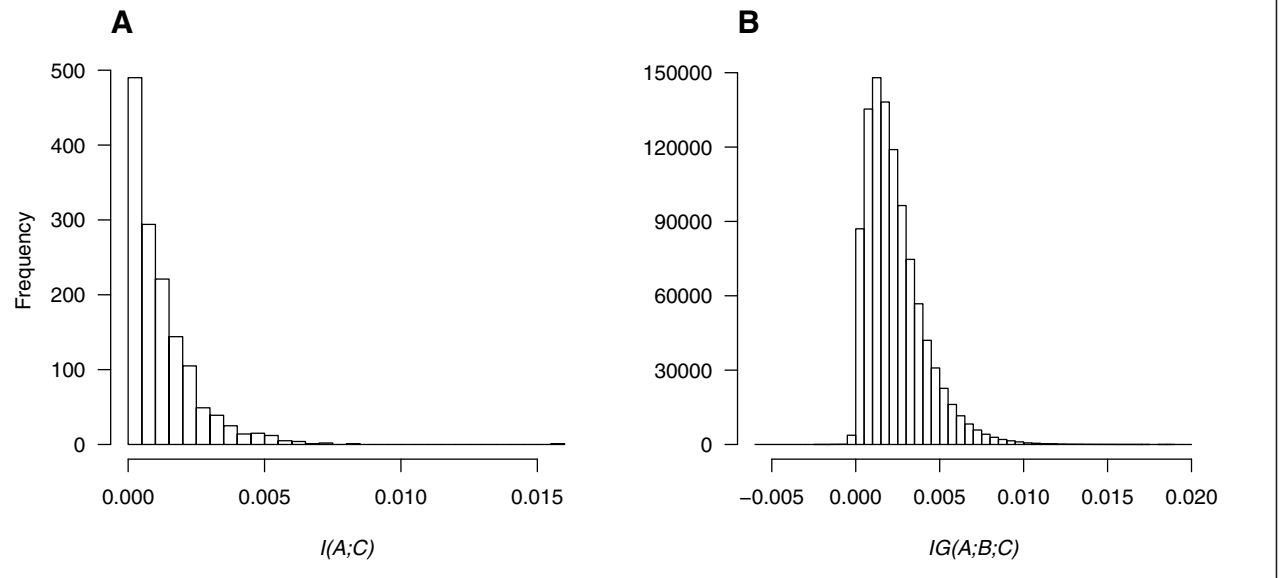

Figure 1 Frequency distributions of the mutual information and the information gain from the real data set. **A** Frequency distribution of main effects for all 1,422 SNPs. The values of $I(A; C)$ range from 0 to 0.01551. **B** Frequency distribution of pairwise interactions for all 1,010,331 pairs of SNPs. The values of $IG(A; B; C)$ range from -0.00591 to 0.01967.

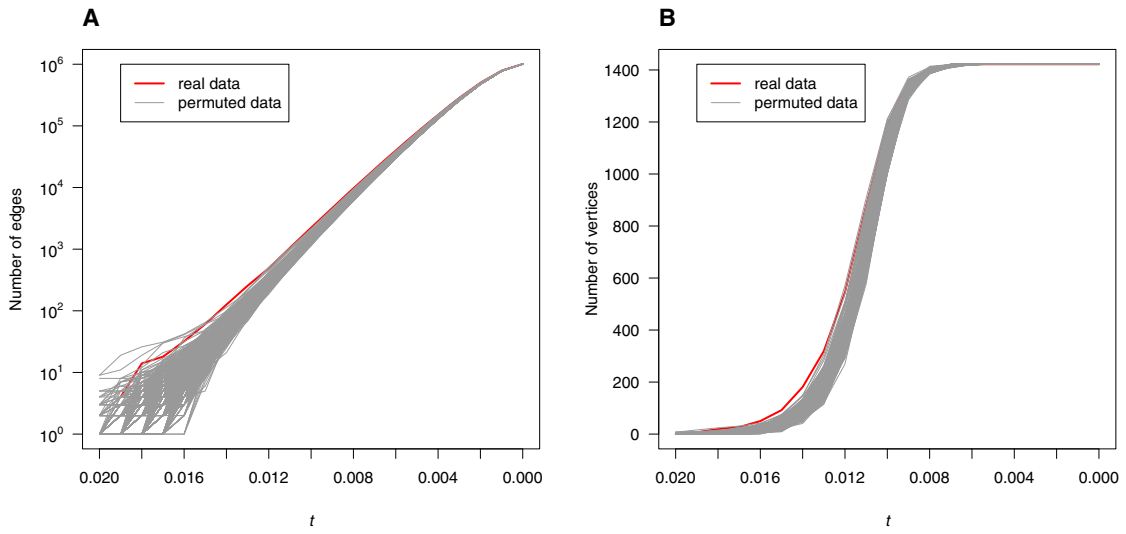

Figure 2 Network growth with decreasing threshold *t*. **A** Increase in the number of edges. **B** Increase in the number of vertices. In both graphs, the red line represents G_t of the real data and the gray lines represent networks of 1,000 permuted data sets. The threshold t decreases from 0.02 to 0 in increments of 0.001.

One might speculate that those observations were not surprising since, for a fixed value of the threshold t, G_t had more edges than did, on average, the graphs constructed from the permuted data (Figure 2).

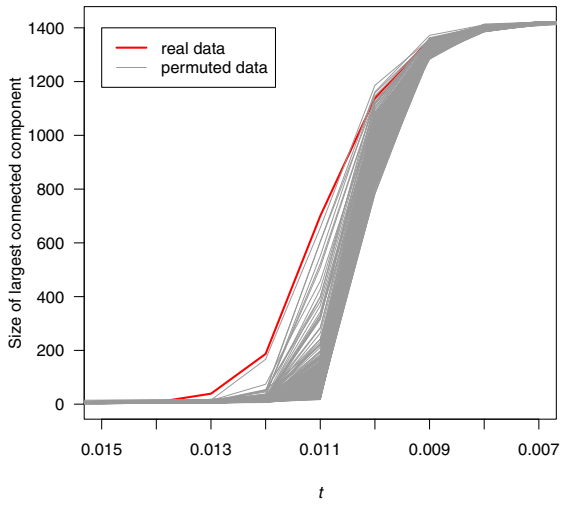

Figure 3 The size of the largest connected component in the networks with decreasing threshold *t*. The red line represents the real-data network G_t and the gray lines represent the networks of 1,000 permuted data sets. The largest connected components include increasingly more vertices as t decreases and eventually include all 1,422 vertices.

However, networks of more edges and vertices do not necessarily have larger and faster growing connected components. The size of the largest connected component essentially characterizes to which extend the vertices of a network are connected to each other. In fact, even for comparable numbers of edges, the differences in growth between the largest connected components of both G_t and the permuted-data graphs persisted. For example, in the real-data graph, an increase in the number of edges of G_t from 255 to 490, as the threshold decreased from 0.013 to 0.012, was accompanied by an increase in the size of the largest connected component of 148, from 39 to 187. In the permuted-data graphs on the other hand, the size of the largest connected component grew only by 54, from 14 to 68, for an increase in edge number of 335 from 270 to 605 as the threshold decreased from 0.012 to 0.011. Thus, both the size of the largest connected component and the rate at which it grew distinguished the G_t from the networks constructed from the permuted data. Based on above observations, t = 0.013 emerged as a threshold of particular interest.

Comparison of vertex degree distributions for the threshold 0.013

Table 1 shows the degree distribution of the network $G_{0.013}$ and of the 1,000 networks constructed from the permuted data using the same value of t. Permuted-data networks had, on average, more vertices with degree one and fewer vertices of higher degrees. In particular, $p(d)$ for the real-data networks always lay

Table 1 Vertex degree distribution of networks for real versus permuted data

d	*p*(*d*)	
	Real Data Set	Permuted Data Sets (*mean* ± *stdev*)
1	0.677	[0.747, 0.831]
2	0.201	[0.119. 0.186]
3	0.0533	[0.0199, 0.0528]
4	0.0345	[0.00184, 0.0210]
5	0.0125	[-0.00168, 0.0124]
6	1.25×10^{-2}	[-1.85×10^{-3}, 5.72×10^{-3}]
7	0	[-1.73×10^{-3}, 5.81×10^{-3}]
8	6.27×10^{-3}	[-1.62×10^{-3}, 3.41×10^{-3}]
9	0	[-1.07×10^{-3}, 1.65×10^{-3}]
10	0	[-8.54×10^{-4}, 1.09×10^{-3}]
11	3.13×10^{-3}	[-3.14×10^{-4}, 3.42×10^{-4}]

The network from the real data has significantly fewer vertices with degree 1 than the networks from the permuted data sets, but more vertices with high degrees.

more than one standard deviation away from the mean of $p(d)$ for the permuted-data networks, except for the three degrees for which the real-data networks had no vertices. This unexpected bias toward high-degree vertices in $G_{0.013}$ led us to consider its degree distribution in more detail and to compare it with the degree distributions of other real-data networks obtained by varying t.

Vertex degree distributions of $\hat{G}_t$

To lessen the risk of including edges likely to exist mostly by chance in G_t, we used $\hat{G}_t$, the subgraph of G_t including only edges with significance $p \leq 0.01$. This changed nothing for $t = 0.013$, as the edges of $G_{0.013}$ all had significance $p \leq 0.001$, but resulted in filtering out edges for lower thresholds.

Figure 4 illustrates part of the vertex degree distributions of the networks $\hat{G}_t$ for $0.013 \geq t \geq 0.006$, i.e., only the points $(d, p(d))$ with $p(d) \neq 0$. Logarithmic scales are used on both axes, so that only points corresponding to nonzero-vertex degrees can be shown. The networks constructed using threshold $t \geq 0.014$ had very few vertices overall and none with degree > 5, and the networks constructed using $t \leq 0.005$ showed very similar patterns to those observed for $t = 0.006$. Therefore, we did not show the degree distributions of these networks.

The vertex degree distributions of $\hat{G}_t$ with $t = 0.013$, 0.012 and 0.011 were approximately linear (Figure 4-A). Since the scale of Figure 4 is logarithmic, these degree distributions can be approximated by functions of the form $p(d) = c \times d^{-\gamma}$ for suitable positive constants c and γ. The graphs of such functions are referred to as power curves. We used *least squares* to find the power curves that best fit the points $(d, p(d))$ for d varying from 1 to the highest nonzero-vertex degree of $\hat{G}_t$. The values of γ

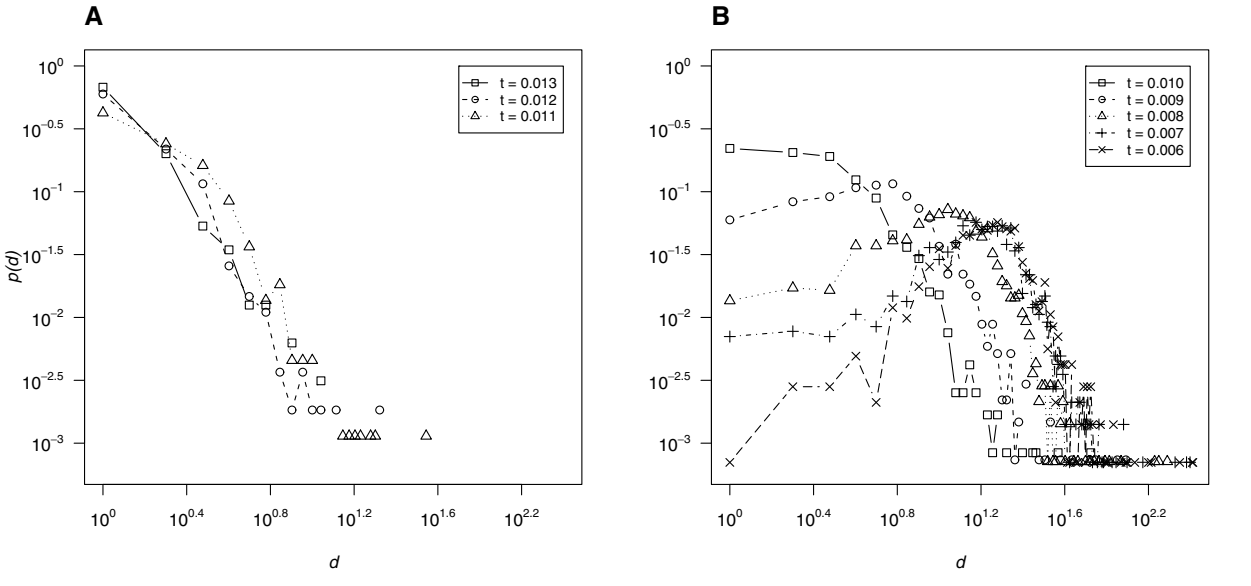

Figure 4 Vertex degree distributions of networks $\hat{G}_t$ with *t* ranging from 0.013 to 0.011 (panel A) and from 0.01 to 0.006 (panel B). Both axes are on logarithmic scale. Each point represents one vertex degree value. Points are only connected by lines if their degree values differ by 1. **A** Approximately linear curves for $0.013 \geq t \geq 0.011$. **B** From 0.01 to 0.006, vertex degree distributions become increasingly bell-shaped.

we found for t = 0.013, 0.012, and 0.011 were 2.01, 1.73, and 1.3, respectively. However, according to the Kolmogorov-Smirnov test, the resulting functions t the degree distributions of $\hat{G}_{0.012}$ and $\hat{G}_{0.011}$ very poorly: for both networks, the null hypothesis that the observed degree distribution follows the best-fitting power curve was rejected with $p < 0.0005$. For $\hat{G}_{0.013}$ on the other hand, the corresponding p value was 0.366, suggesting that the null hypothesis was still plausible. Figure 5 shows the degree distribution of $\hat{G}_{0.013}$ and the fitting power curve for $p(d) = 0.615 \times d^{-2.01}$.

The vertex degree distributions of $\hat{G}_t$ became increasingly bell-shaped as t decreased from 0.010 to 0.006 (Figure 4-B). This occurred as more edges of low weight were likely to be included in $\hat{G}_t$ due to chance rather than to biological significance and $\hat{G}_t$ therefore progressively resembled random networks. The vertex degree distributions of such random networks follow a Poisson distribution $p(d) = \frac{\lambda^d}{d!}e^{-\lambda}$, where λ is the mean, in our case the average vertex degree (see Additional file 1 for the Poisson distribution fitting curves for each cutoff t).

A vertex degree distribution can still follow a Poisson distribution even if it is not bell-shaped, which happens when λ is small. For $\hat{G}_{0.013}$, $\lambda = \frac{2\times255}{1,422} \approx 0.366$, where 255 was the number of edges in $\hat{G}_{0.013}$ and 1,422 was the total number of SNPs. For such a small value of λ, a Poisson distribution is not bell-shaped. Hence, ruling out the possibility that $\hat{G}_{0.013}$ follows a Poisson degree distribution required further investigation.

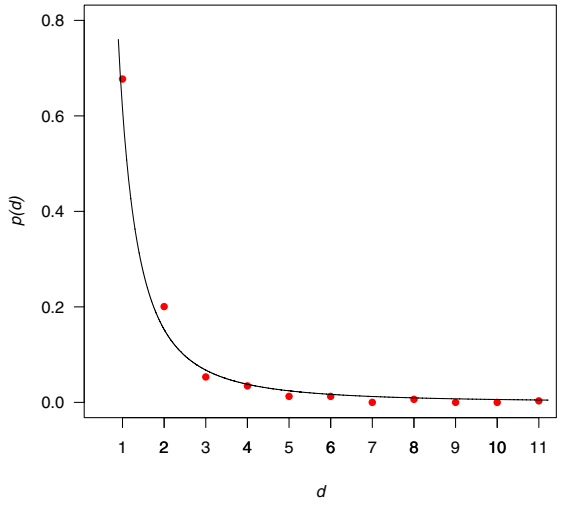

Figure 5 Vertex degree distribution of network $\hat{G}_{0.013}$. The red points show the observed values and the line in black is the fitting power-law curve of $p(d) = 0.615 \times d^{-2.01}$.

We therefore tested the hypothesis that the vertex degrees of $\hat{G}_{0.013}$ followed a Poisson distribution. The construction process of the networks $\hat{G}_t$ can be described as attaching edges to 1,422 vertices and then removing the vertices of degree zero. If this attachment were random, and no degree-zero vertices were removed, the vertex degrees would follow a Poisson distribution. When degree-zero vertices are removed, as was the case here, the theoretical Poisson distribution has to be adjusted as follows:

$$P_0(d) = \begin{cases} \frac{\lambda^d}{kd!}e^{-\lambda} & \text{if } d \geq 1 \\ P_0(0) = 0 \end{cases} \quad (5)$$

where $k = 1 - P(0) = 1 - e^{-\lambda}$ normalized the adjusted distribution $P_0(d)$ since $P_0(0) = 0$. According to the Kolmogorov-Smirnov test, the null hypothesis that the vertex degrees of $\hat{G}_{0.013}$ were drawn from the adjusted Poisson distribution $P_0(d)$ or, equivalently, that its edge attachment was random was rejected with $p = 0.001$.

Networks with degree distributions of the form $p(d) = c \times d^{-\lambda}$ are said to have power-law distributions and are often called scale-free in the literature [30]. Although the term is usually only applied to very large networks, at least two to three magnitudes larger than those considered here, our results nevertheless suggest that the network $G_{0.013}$ was scale-free, or at least approximately so.

Network $G_{0.013}$

The network $G_{0.013}$ (Figure 6) had 255 edges, 319 vertices, and 79 connected components (see Additional files 2, 3, 4 for subdivided graphs with only the largest component, other relatively large components, and the rest small components). All of those 255 edges have significance $p \leq 0.001$. This could be partially explained by the fact that these top 255 edges had relatively high weights and thus more likely obtained smaller p-values using permutation testing. The largest connected component had 39 vertices. This was more than twice as large as the second largest connected component of size 18. In Figure 6, the size of a vertex is proportional to the main effect of the corresponding SNP and the width of an edge is proportional to the strength of the interaction between the two SNPs it joins (see Additional files 5 and 6 for the standard network vertex and edge files). The network provides a clear visualization of the pairs of SNPs which had the strongest synergetic effect on bladder cancer, as well as the strength of these effects and of the individual SNPs involved in the strongest interactions. Most importantly, the network shows which synergetic pairs

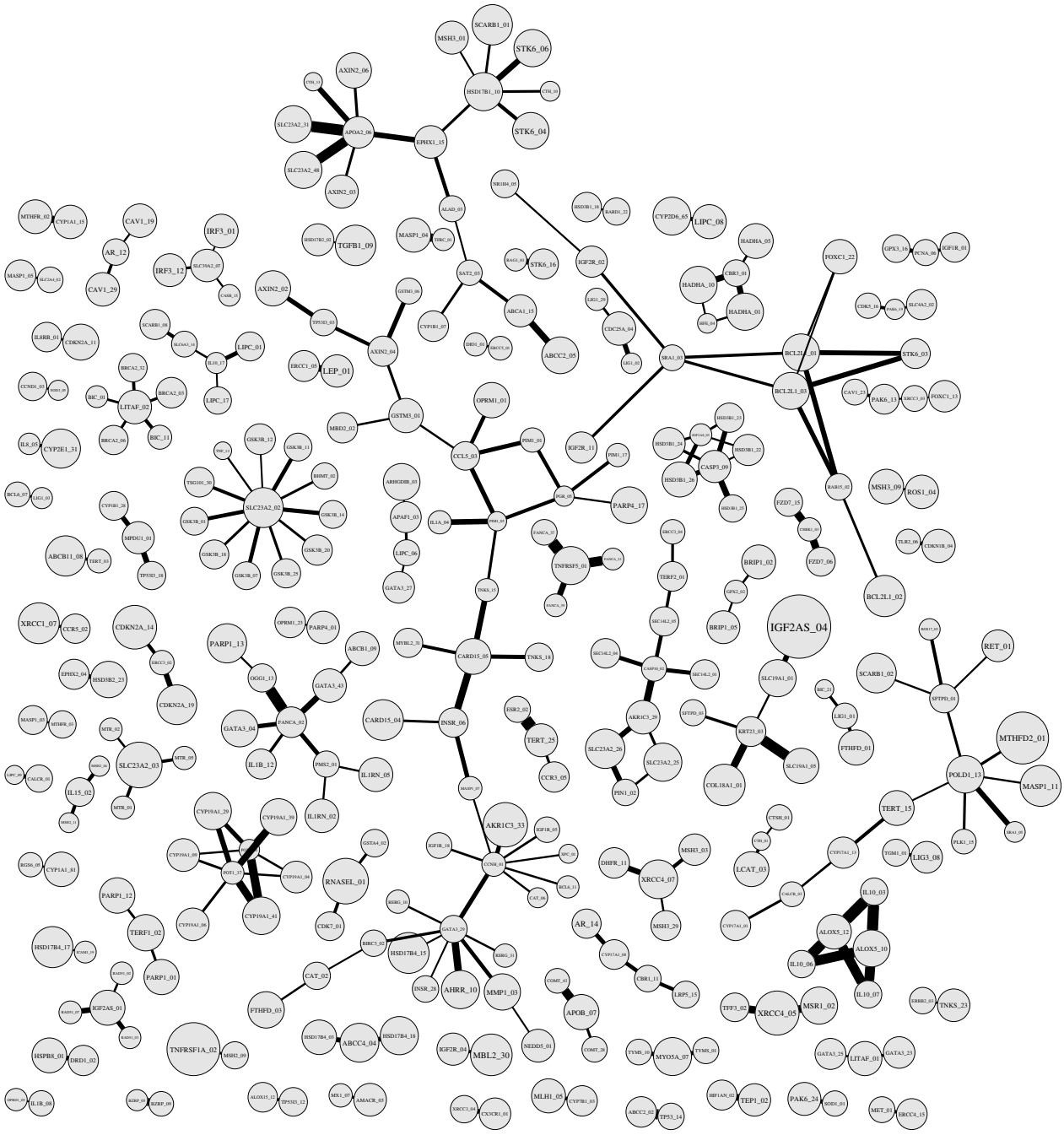

Figure 6 Statistical epistasis network $\hat{G}_{0.013}$. There are 319 vertices and 255 edges. The network has 79 connected components and the largest one has 39 vertices. The width of an edge and the size of a vertex are in proportion to their weights. The length of an edge is for layout purposes only. The graph is rendered by the software Graphviz.

shared a SNP, and thereby captures the entire pairwise interaction space.

As is the case for biological pathways, this statistical epistasis network showed very few cycles. In particular, there were no connected triangles. That is, vertices did not interact with their neighbors' neighbors. Moreover, in accordance with its power-law degree distribution, the network had a few vertices with degrees that were much higher than the average, while the majority of vertices connected directly to only one other vertex. Finally, vertices with high degrees or connected with wide edges were not necessarily of large size (see Additional file 7

for the linear regression showing no correlation between vertex size and vertex connection).

Discussion

The goal of this study was to infer and characterize statistical epistasis networks in a large population-based study of bladder cancer susceptibility. We observed distinguishing topologies of the networks assembled using the cancer data and the implication that a group of SNPs may jointly modify the disease outcome. Specifically, the networks G_t had many more high-degree vertices and their largest connected components emerged earlier and grew faster than expected. These characteristics were the most apparent when $t = 0.013$. The network $G_{0.013}$ was shown to be approximately scale-free, an important property found in various natural and social networks. This property was no longer observable when t further decreased and edges representing weaker and possibly less biologically relevant pairwise interactions were added.

The network $G_{0.013}$ allows for some interesting observations about the structure of the pairwise interaction space of the genetic data. First, SNPs aggregate to form connected components, which may indicate that multiple SNPs jointly modify disease outcome. In $G_{0.013}$, SNPs are grouped into 79 connected components of size ranging from 2 to 39. These connected components show various structural patterns, also known as *motifs*, including lines, crosses, and stars. The largest connected component has a tree-like structure. This may imply the existence of unique interaction patterns among groups of SNPs.

Second, the network has an approximately scale-free topology and an ensemble of particularly high-degree vertices, which suggests that it may be exceptionally robust. Scale-free networks permeate natural and social sciences [47-49]. The most well-known scale-free networks are the backbone of the Internet and social networks. In biology, scale-free topologies have been found in metabolic networks [31], protein-protein interaction networks [33], and gene-regulatory networks [34]. Those various scale-free networks share an intriguing property: the value of γ in the degree distributions $p(d) = c \times d^{-\gamma}$ mostly satisfies $2 \le \gamma \le 3$ [47], which is also the case for $G_{0.013}$ ($\gamma = 2.01$). As more scale-free networks are being discovered in a variety of fields, a question remains: how can systems as fundamentally different as the cell and the Internet have a similar architecture and obey the same laws [47]? Scale-free networks typically have many vertices with low degrees and a few vertices with high degrees, also known as *hubs* [30]. This essentially differentiates scale-free networks from random networks where the majority of vertices have average degrees. The probability $p(d)$ of degree d in the Poisson distribution decreases exponentially as d increases, and thus random networks are very unlikely to have hubs with degrees much larger than the average. The existence of hubs in a scale-free network implies strong robustness against failures. Because random vertex removal is very unlikely to affect hubs, the connectivity of the network most likely remains intact. In biological networks, this robustness translates into the resilience of organisms to intrinsic and environmental perturbations. For instance, in protein-protein interaction networks [33], most proteins interact with only one or two other proteins but a few are able to interact to a large number. Such hub proteins are rarely affected by mutations and organisms can remain functional under most perturbations. The simultaneous emergence of scale-free topologies in many biological networks suggests that evolution has favored such a structure in natural systems. Moreover, it suggests that the robustness of natural systems does not only result from inherent genetic redundancy but also, and maybe more importantly, from the topological organization of entities and interactions [33]. Although our epistasis network is developed based on statistical rather than on real bio-chemical interactions, it is interesting to observe similar topologies between biological and statistical networks.

Third, the existence of main effects does not necessarily correlate with the occurrence of interactions. This, in turn, suggests that many current main-effect-prioritized methods might have overlooked SNPs contributing to the disease susceptibility through their interactions with other SNPs rather than through their main effects. As shown in the graph, large main-effect SNPs do not necessarily associate with strong pairwise interactions or interact with many other SNPs. Instead, SNPs involved in potential important pairwise interactions, such as those located on the central path of the largest connected component, often have relatively small main effects.

The statistical epistasis network approach has many advantages. 1) Networks allow for efficiently visualizing both main and epistatic effects and how they interplay. 2) Networks serve as a very intuitive tool to study pairwise interactions and to characterize the entire epistatic interaction space. Moreover, they may also help identify higher-order interactions by grouping SNPs into connected components. High-order epistasis does not necessarily require detectable pairwise interactions between SNPs. However, given that current computational power allows only for exhaustively enumerating pairwise interactions in moderate-size data sets, pairwise interaction networks may serve as a useful guide to explore higher-order epistasis among SNPs that exhibit lower-order interactions. 3) Our network model is assembled using the entire set of available SNPs without

limiting ourselves to only high main-effect ones. This reduces the risk of overlooking candidate SNPs that are involved in important interactions but with low main effects. 4) Network topological analyses are used to systematically determine the best network that captures the genetic architecture of a data set. 5) Networks, along with graph theory, are well-developed fields, and many advanced techniques and analytical tools are likely to benefit future network-based epistasis studies. In particular, additional topological properties such as motif distribution and network diameter [30,42] are worth investigating.

Among the limitations of this approach is that it is still under development and no user-friendly interface is available yet. Different data sets may require different analytical tools and a fully automated analysis software may therefore not be appropriate. Moreover, since the approach aims at highlighting pairs of SNPs with strong pairwise interactions, it is likely to overlook SNPs that are only involved in higher-order interactions. As mentioned previously, strong three- or higher-order interactions may exist despite the absence of pairwise interactions.

The statistical epistasis network approach we used can be extended in the following ways. 1) The network $G_{0.013}$ will be further studied for bladder cancer association. Through a closer investigation, such as gene ontologies and biological pathways, on those 319 SNPs in the network, especially those 39 SNPs in the largest connected component, we expect to prioritize gene categories with high bladder cancer susceptibility, and to testify whether SNP interactions tend to happen within the same category or across categories. Other possible applications include using the network to train classifiers in predicting bladder cancer risk [50] and to supervise data mining methods for identifying high-order genetic interactions [27]. 2) The approach can also be applied to other data sets. We are particularly interested in investigating network topologies in larger data sets or data associated with other diseases. 3) To corroborate the present results, future studies could use metrics other than information theoretical measures, such as SNP and gene annotation or SURF scores, which are obtained by directly assessing genetic variants depending on their phenotype relevance using machine learning techniques [51]. 4) Given the effect of smoking [37] and arsenic exposure [41,52] on bladder cancer prevalence, an additional next step is to account for gene-environment interactions in our analyses. This can be achieved by adding these environmental factors to our model, and investigating how the environmental background on which the genes are expressed modify the conclusions we draw.

Conclusions

In this study, we proposed a statistical epistasis network approach that is able to capture the global landscape of gene-gene interactions in a large population-based bladder cancer data set. Through an exhaustive enumeration of all possible pairwise interactions and network topological analyses, a distinctive network is systematically identified which shows unique properties. It has a significantly large connected component and an intriguing approximate scale-free topology that permeate natural and technical networks. Specifically in the context of biological networks, scale-free is well recognized as an outcome interaction topology of robust organisms resulted by natural evolution. The observation of such a network topology in the bladder cancer data set may indicate a global interactive structure embedded in the genetic architecture of bladder cancer.

The derived network from this study may further benefit bladder cancer studies through closer examinations of SNP characteristics. In addition to a global interaction picture of bladder cancer depicted by this network, further studies on individual gene ontology and biological pathway categorization may provide important insight on prioritizing inter- or intra-category genetic interactions. Moreover, the proposed network approach holds the promise characterizing a broader gene-gene interaction landscape in epistasis studies, and is expected to be applied to other data sets, especially large-scale ones.

Additional material

Additional file 1: Poisson vertex degree distribution fitting curves of networks $\hat{G}_t$ with *t* ranging from 0.013 to 0.011 (panel A) and from 0.01 to 0.006 (panel B). If networks $\hat{G}_t$ were built through the process of randomly linking two vertices and then removing degree-zero vertices, their vertex degrees would follow an adjusted Poisson distribution $P_0(d) = \frac{\lambda^d}{kd!}e^{-\lambda}$, $d > 0$, where the normalizing factor $k = P(0) = 1 - e^{-\lambda}$ and λ is the average vertex degree of networks $\hat{G}_t$. Both axes are on logarithmic scale.

Additional file 2: The largest connected component in network $\hat{G}_{0.013}$. There are 39 SNPs connected in the largest component.

Additional file 3: Other large connected components in network $\hat{G}_{0.013}$. The sizes of the other large connected components are ranging from 5 to 18.

Additional file 4: Small connected components in network $\hat{G}_{0.013}$. The small connected components only have 2 to 4 SNPs.

Additional file 5: Standard network vertex file of $\hat{G}_{0.013}$. The file shows a list of vertices and their weights.

Additional file 6: Standard network edge file of $\hat{G}_{0.013}$. The file shows a list of edges and their weights.

Additional file 7: Vertex main effect as a function of degree (panel A) and the total weight of attached edges (panel B) in network $\hat{G}_{0.013}$. The vertex main effect is independent of its degree and summed weight of all attached edges. Lines show the correlations using linear regression.

Acknowledgements and funding
We would like to thank Davnah Urbach for her editorial help. This work was supported by the National Institutes of Health R01-LM009012, R01-LM010098, and R01-AI59694 to JHM, R01-CA57494 and P42-ES007373 to MRK, K07-CA102327 and R03-CA121382 to ASA.

Author details
[1]Department of Genetics, Dartmouth Medical School, Dartmouth College, Lebanon, NH, USA. [2]Department of Community and Family Medicine, Dartmouth Medical School, Dartmouth College, Lebanon, NH, USA. [3]Institute for Quantitative Biomedical Sciences, Dartmouth Medical School, Dartmouth College, Lebanon, NH, USA.

Authors' contributions
TH designed the study, performed the analyses, and drafted the manuscript. NASA participated in the design of the study and performed the analyses. JWK participated in the analyses and helped to draft the manuscript. ASA and MRK carried out the data collection and the genotyping, and helped to draft the manuscript. JHM conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 25 April 2011 Accepted: 12 September 2011
Published: 12 September 2011

References

1. Merikangas KR, Low NCP, Hardy J: **Commentary: Understanding sources of complexity in chronic diseases - the importance of integration of genetics and epidemiology.** *International Journal of Epidemiology* 2006, **35**:590-592.
2. Hirschhorn JN, Daly MJ: **Genome-Wide Association Studies for Common Diseases and Complex Traits.** *Nature Review Genetics* 2005, **6(2)**:95-108.
3. Hirschhorn JN: **Genomewide Association Studies – Illuminating Biologic Pathways.** *The New England Journal of Medicine* 2009, **360(17)**:1699-1701.
4. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Research* 2009, **37**:D5-D15.
5. Crawford DC, Dilks HH: **Strategies for Genotyping.** *Current Protocols in Human Genetics* 2011, **1**:Unit 1.3.
6. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP: **Computational solutions to large-scale data management and analysis.** *Nature Review Genetics* 2010, **11**:647-657.
7. Wang WYS, Barratt BJ, Clayton DG, Todd JA: **Genome-Wide Association Studies: Theoretical and Practical Concerns.** *Nature Review Genetics* 2005, **6(2)**:109-118.
8. Hardy J, Singleton A: **Genome-Wide Association Studies and Human Disease.** *New England Journal of Medicine* 2009, **360(17)**:1759-1768.
9. Hindorff LA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proceedings of the National Academy of Sciences* 2009, **106(23)**:9362-9367.
10. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.
11. Moore JH, Ritchie MD: **The challenges of whole-genome approaches to common diseases.** *Journal of the Amarican Medical Association* 2004, **291(13)**:1642-1643.
12. Clark AG, Boerwinkle E, Hixson J, Sing CF: **Determinants of the success of whole-genome association testing.** *Genome Research* 2005, **15**:1463-1467.
13. Moore JH, Williams SM: **Epistasis and Its Implications for Personal Genetics.** *The American Journal of Human Genetics* 2009, **85(3)**:309-320.
14. Phillips PC: **The Language of Gene Interaction.** *Genetics* 1998, **149**:1167-1171.
15. Templeton AR: **Epistasis and Complex Traits.** In *Epistasis and the Evolutionary Process.* Edited by: Wolf JB, Brodie ED, Wade MJ. Oxford University Press; 2000:41-57.
16. Cordell HJ: **Epistasis: What It Means, What It Doesn't Mean, and Statistical Methods to Detect It in Humans.** *Human Molecular Genetics* 2002, **11(20)**:2463-2468.
17. Moore JH, Williams SM: **Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis.** *BioEssays* 2005, **27(6)**:637-646.
18. Phillips PC: **Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems.** *Nature Review Genetics* 2008, **9**:855-867.
19. Tyler AL, Asselbergs FW, Williams SM, Moore JH: **Shadows of complexity: what biological networks reveal about epistasis and pleiotropy.** *BioEssays* 2009, **31**:220-227.
20. Musani SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari HK, Allison DB: **Detection of Gene Gene Interactions in Genome-Wide Association Studies of Human Population Data.** *Human Heredity* 2007, **63**:67-84.
21. Cordell HJ: **Detecting gene-gene interactions that underlie human diseases.** *Nature Review Genetics* 2009, **10(6)**:392-404.
22. Moore JH, Asselbergs FW, Williams SM: **Bioinformatics challenges for genome-wide association studies.** *Bioinformatics* 2010, **26(4)**:445-455.
23. Thornton-Wells TA, Moore JH, Haines JL: **Genetics, statistics and human disease: analytical retooling for complexity.** *Trends in Genetics* 2004, **20(12)**:640-647.
24. Moore JH, White BC: **Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge.** In *Genetic Programming Theory and Practice IV.* Edited by: Riolo RL, Soule T, Worzel B. Springer; 2005:969-977.
25. Eppstein MJ, Payne JL, White BC, Moore JH: **Genomic Mining For Complex Disease Traits with 'Random Chemistry'.** *Genetic Programming and Evolvable Machines* 2007, **8(4)**:395-411.
26. Greene CS, Moore JH: **Solving complex problems in human genetics using nature-inspired algorithms: Strategies for exploiting domain-specific knowledge.** In *Nature Inspired Informatics.* Edited by: Chiong R. IGI Global; 2009:166-180.
27. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: **Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer.** *The American Journal of Human Genetics* 2001, **69**:138-147.
28. Hahn LW, Ritchie MD, Moore JH: **Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions.** *Bioinformatics* 2003, **19(3)**:376-382.
29. Strogatz SH: **Exploring complex networks.** *Nature* 2001, **410**:268-276.
30. Newman MEJ: *Networks: An Introduction* Oxford University Press; 2010.
31. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407**:651-654.
32. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297**:1551-1555.
33. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411**:41-42.
34. Barabasi AL, Oltvai ZN: **Network biology: Understanding the cell's functional organization.** *Nature Review Genetics* 2004, **5**:101-113.
35. Martinez ND: **Constant Connectance in Community Food Webs.** *The American Society of Naturalists* 1992, **140(6)**:1208-1218.
36. McKinney BA, Crowe JE, Guo J, Tian D: **Capturing the Spectrum of Interaction Effects in Genetic Association Studies by Simulated Evaporative Cooling Network Analysis.** *PLoS Genetics* 2009, **5(3)**:e1000432.
37. Silverman DT, Morrison AS, Devesa SS: **Bladder Cancer.** In *Cancer Epidemiology and Prevention.* Edited by: Schottenfeld D, Fraumeni JFJ. New York, NY, USA: Oxford University Press; 1996:1156-1179.
38. Karagas MR, Park S, Nelson HH, Andrew AS, Mott L, Schned A, Kelsey KT: **Methylenetetrahydrofolate reductase (MTHFR) variants and bladder cancer: a population-based case-control study.** *International Journal of Hygiene and Environmental Health* 2005, **208(5)**:321-327.
39. Garcia-Closas M, Malats N, Silverman D, Dosemeci M, Kogevinas M, *et al*: **NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses.** *The Lancet* 2005, **366(9486)**:649-659.
40. Andrew AS, Nelson HH, Kelsey KT, Moore JH, Meng AC, Casella DP, Tosteson TD, Schned AR, Karagas MR: **Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility.** *Carcinogenesis* 2006, **27(5)**:1030-1037.
41. Karagas MR, Tosteson TD, Blum J, Morris JS, Baron JA, Klaue B: **Design of an epidemiologic study of drinking water arsenic exposure and skin and**

bladder cancer risk in a U.S. population. *Environmental Health Perspectives* 1998, **106(4)**:1047-1050.
42. West DB: *Introduction to Graph Theory*. Second edition. Prentice Hall; 2001.
43. Cover TM, Thomas JA: *Elements of Information Theory*. Second edition. Wiley; 2006.
44. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC: **A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility.** *Journal of Theoretical Biology* 2006, **241(2)**:252-261.
45. Moore JH, Barney N, Tsai CT, Chiang FT, Gui J, White BC: **Symbolic Modeling of Epistasis.** *Human Heredity* 2007, **63(2)**:120-133.
46. Jakulin A, Bratko I: **Analyzing Attribute Dependencies.** *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2003), Volume 2838 of Lecture Notes in Artificial Intelligence* Springer-Verlag; 2003, 229-240.
47. Barabasi AL, Bonabeau E: **Scale-Free Networks.** *Scientific American* 2003, **5**:50-59.
48. Li L, Alderson D, Doyle JC, Willinger W: **Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications.** *Internet Mathematics* 2005, **2(4)**:431-523.
49. Newman MEJ: **Power laws, Pareto distributions and Zipf's law.** *Contemporary Physics* 2005, **46(5)**:323-351.
50. Li X, Rao S, Wang Y, Gong B: **Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling.** *Nucleic Acids Research* 2004, **32(9)**:2685-2694.
51. Greene CS, Penrod N, Kiralis J, Moore JH: **Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions.** *BioData Mining* 2009, **2(5)**.
52. Chen CJ, Chen CW, Wu MM, Kuo TL: **Cancer potential in liver, lung, bladder and kidney due to ingested inorganic arsenic in drinking water.** *British Journal of Cancer* 1992, **66(5)**:888-892.

doi:10.1186/1471-2105-12-364
Cite this article as: Hu *et al.*: Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics* 2011 **12**:364.